APPLICATION

FOR

UNITED STATES LETTERS PATENT

TITLE: UNSUPERVISED LEARNING OF OBJECT CATEGORIES FROM CLUTTERED IMAGES

APPLICANT: PIETRO PERONA, MARKUS WEBER AND MAX WELLING

CERTIFICATE OF MAILING BY EXPRESS MAIL

Express Mail Label No. EL870691579US

I hereby certify that this correspondence is being deposited with the United States Postal Service as Express Mail Post Office to Addressee with sufficient postage on the date indicated below and is addressed to the Commissioner for Patents, Washington, D.C. 20231.

February 1, 2002
Date of Deposit

Signature

Gabriel Lewis
Typed or Printed Name of Person Signing Certificate

UNSUPERVISED LEARNING OF OBJECT CATEGORIES FROM CLUTTERED IMAGES

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] The present application claims priority from provisional application number 60/266,014, filed February 1, 2001.

STATEMENT AS TO FEDERALLY-SPONSORED RESEARCH

[0002] The U.S. Government has certain rights in this invention pursuant to Grant Nos. NSF9402726 and NSF9457618 awarded by the National Science Foundation.

BACKGROUND

[0003] Machines can be used to review an electronic version of an image to recognize information within an electronic image.

[0004] An object class is a collection of objects which share characteristic parts or features that are visually similar, and which occur in similar spatial configurations. Solutions to the problem of recognizing members of object classes have taken various forms. Use of various models have been suggested.

[0005] These techniques often require a training process that attempts to carry out:

-segmentation, that is which objects are to be recognized and where do they appear in the training images,

-selection, that is, selection of which object parts are distinctive and stable, and

-estimation of model parameters, that is, what parameters of the global geometry or shape and the appearance of the individual parts best describe the training data.

[0006] Previous model based techniques may require a supervised stage of learning. For example, targeted objects must be identified in training images either by selection of points or regions on their surfaces, or by segmenting the objects from the background. This may produce significant disadvantages, including that any prejudices of the human observer, such as which features appear most distinctive to the human observer, may also be trained during the training process.

SUMMARY

[0007] The present application defines a statistical model in which shape variability is modeled in a probabilistic setting. In an embodiment, a set of images

are used to automatically train a system for model learning. The model is formed based on probabilistic techniques.

[0008] In an embodiment, an interest operator is used to determine textured regions in the image, and the number of features are reduced by clustering. An automatic model estimates which of these features are the most important and probabilistically determines a model, a correspondence, and a joint model probability density. This may be done using expectation maximization.

BRIEF DESCRIPTION OF THE DRAWINGS

[0009] These and other aspects will now be described in detail with reference to the accompanying drawings, wherein:

[0010] Figure 1 shows a basic block diagram of the system of the present application;

[0011] Figure 2 shows a flowchart of the basic training techniques;

[0012] Figure 3 shows some generic templates;

[0013] Figure 4 shows a flowchart of the overall operation; and

[0014] Figure 5 shows some exemplary results for models for specified letters.

DETAILED DESCRIPTION

[0015] In the present system, instances of an object class are described through characteristic set of features/parts which can occur at varying spatial locations. The objects are composed of parts and shapes. Parts are the image patches that may be detected and characterized by detectors. Shape describes the geometry of the parts. In the embodiment, a joint probability density is based on part appearance and shape models of the object class.

[0016] The parts are modeled as rigid patterns. Their positional variability is represented using a probability density function over the point locations of the object features. Translation of the part features are eliminated by describing all feature positions relative to one reference feature. Positions are represented by a Gaussian probability density function.

[0017] In one aspect, the system determines whether the image contains only clutter or "background", or whether the image contains an instance of the class or "foreground". According to an embodiment, the object features may be independently detected using different types of feature

detectors. After detecting the features, a hypothesis evaluation stage evaluates candidate locations in the image to determine the likelihood of their actually corresponding to an instance of the object class. This is done by fitting a mixture density model to the data. The mixture density model includes a joint Gaussian density over all foreground detector responses, and a uniform density over background responses.

[0018] Figure 1 shows a basic diagram of the operation and Figure 4 shows a flowchart of the basic operation. In general, the operation can be carried out on any programmed computer. Gigure 1 may be embodied in a general-purpose computer, as software within the computer, or as any kind of hardware system including dedicated logic, programmable logic, and the like. The images may be obtained from files, or may be obtained using a camera.

[0019] An image set 100 may be used for automatic feature selection at 400. The image set is applied to a feature selection system 110. The feature selection system 110 may have an interest operator 112 which automatically detects textured regions in the images. The interest operator may be the so called Forstner interest operator. This interest operator may detect corner points, line intersections, center points and the like. Figure 3 shows

a set of 14 generic detector templates which may be used. These templates are normalized such that their mean is equal to zero. Other techniques may be used, however, using other templates.

[0020] The automatic feature selection in 400 may produce 10,000 or more features per image.

[0021] The number of interesting features is reduced in a vector quantizer 114 which quantizes the vectors and clusters them by grouping similar parts. A clustering algorithm may also be used. This may produce approximately 150 features per image. Shifting by multiple pixels may further reduce the redundancy.

[0022] The object model is trained using the part candidates.

[0023] Model training at 410 trains the feature detectors using the resultant clusters, in the model learning block 120. This is done to estimate which are of the features are actually the most informative, and to determine the probabilistic description of the constellation that these features form when they are exposed to an object of interest. This is done by forming the model structure, establishing a correspondence between homologous parts across the training set, and labeling and other parts as background or noise.

[0024] The inventors recognize three basic issues which may produce advantages over the prior art. First, the technique used for training should be automated, that is, it should avoid segmentation or labeling of the images manually. Second, a large number of feature detectors should be used to enable selecting certain feature detectors that can consistently identify a shared feature of the object class. This means that a subset of the feature detectors may be selected to choose the model configuration. A global shape representation should also be learned autonomously.

[0025] Training a model requires determining the key parts of the object, selecting corresponding parts on the number of training images, and estimating the joint probability function based on part appearance and shape. While previous practitioners have done this manually, the present technique may automate this.

[0026] The operation proceeds according to the flowchart of figure 2. Initially, a number of feature detectors F may be selected to be part of the model. At 200, all the information is extracted from the training image. The objects are modeled as collections of rigid parts. Each of those parts is detected by a detector, thereby transforming the entire image into a collection of parts. Some of those

parts will correspond to the foreground, that is they will be an instance of the target object class. Other parts stem from background clutter or false detections known as the background.

[0027] Assume T different types of parts. Then, the positions of all parts extracted from 1 image may be summarized as a matrix of feature candidate positions of the form:

$$X^o = \begin{pmatrix} x_{11}x_{12},\ldots,x_{1N_1} \\ x_{21}x_{22},\ldots,x_{2N_2} \\ \vdots \\ x_{T1}x_{T2},\ldots,x_{TN_T} \end{pmatrix},$$

[0028] Each row contains the two-dimensional locations of detections of the feature type F. Random variables of the type

$$\mathcal{D} = \{X_\tau^o, x_\tau^m, \mathbf{n}_\tau, \mathbf{h}_\tau, \mathbf{b}_\tau\}.$$

may be used to represent the explicit or unobserved information. The superscript "o" indicates that the positions are observed, while unobserved features are designated by the superscript "m" for missing.

[0029] The entire set X of feature candidates can be divided between candidates which are be true features of the object or the "foreground", and noise features also called the "background". The random variable vector **h** may be used to create a set of indices so that if $\mathbf{h}_i = j;, j>0$, if the point x_{ij} is a foreground point. If an object part is not included in X^0 then the corresponding entry in **h** will be zero.

[0030] When presented with an unlabeled image, the system does not know which parts correspond to the foreground. This means that **h** is not observable. Therefore, **h** is a hypothesis, since it is used to hypothesize that certain parts of X^0 belong to the foreground object. Positions of the occluded or missed foreground features are collected in a separate vector $\mathbf{x}^m$, where the size of $\mathbf{x}^m$ varies between 0 and F depending on the number of unobserved features. The binary vector b encodes information about which parts have been detected and which omissions or occluded. Therefore, **b**f is 1 if **h**f>0 (the object part is included in X^0), and is 0 otherwise.

[0031] The vector N denotes the number of background candidates included in a specific row of X^0.

[0032] Finally, the number n_{tau} represents the number of background detections.

[0033] At 210, the statistics of the training image is assessed. The object is to classify the images into the classes of whether the object is present (c1) or whether the object is absent (c0). This may be done by choosing the class with the maximum a posteriori probability. The techniques are disclosed herein. This classification may be characterized by the ratio

$$\frac{p(C_1|X^o)}{p(C_0|X^o)} \propto \frac{\sum_{\mathbf{h}} p(X^o, \mathbf{h}|C_1)}{p(X^o, \mathbf{h}_0|C_0)},$$

[0034] The probability distribution modeling the data may be shown as

$$p(X^o, \mathbf{x}^m, \mathbf{h}, \mathbf{n}, \mathbf{b}) = p(X^o, \mathbf{x}^m|\mathbf{h}, \mathbf{n})\, p(\mathbf{h}|\mathbf{n}, \mathbf{b})\, p(\mathbf{n})\, p(\mathbf{b}).$$

[0035] The probability density over the number of background detections may be modeled by a Poisson distribution as

$$p(\mathbf{n}) = \prod_{f=1}^{F} \frac{1}{n_f!}(M_f)^{n_f} e^{-M_f},$$

Where M_f is the average number of background detections per image. Allowing a different M_f for each feature allows modeling different detector statistics and ultimately

enables distinguishing between more reliable detectors and less reliable detectors.

[0036] The vector **b** encodes information about which features have been detected and which are missed. The probability that **b** is 1, p(**b**), is modeled by a table of size 2^F which equals the number of possible binary vectors of length F. If F is large, then the explicit probability mass table of length 2^F may become even longer. Independence between the feature detectors and the model p(**b**) is shown as:

$$p(\mathbf{b}) = \prod_{f=1}^{F} p(b_f).$$

The number of parameters reduces in that case from 2^F to F.

[0037] The density p is modeled by

$$p(\mathbf{h}|\mathbf{n}, \mathbf{b}) = \begin{cases} \frac{1}{\prod_{f=1}^{F} N_f^{b_f}} & \mathbf{h} \in \mathcal{H}_\mathbf{b} \\ 0 & \text{other } \mathbf{h} \end{cases}$$

where Hb denotes the set of all hypotheses consistent with both **b** and **n** and N_f and denotes the total number of detections of the feature f.

[0038] The hypothesized foreground detections are shown as

$$p(X^o, \mathbf{x}^m|\mathbf{h}, \mathbf{n}) = G(\mathbf{z}|\mu, \Sigma)\, U(\mathbf{x}_{bg}),$$

Where $\mathbf{x}^T = (\mathbf{x}^o \mathbf{x}^m)$ is defined as the coordinates of the hypothesized foreground detections both observed and missing, x_{bg} is defined as the coordinates of the background detection, $G(\bar{z}|\mu, \Sigma)$ denotes a Gaussian with a mean of μ and covariance of Σ.

[0039] The positions of the background detections are modified with a uniform intensity shown by

$$U(\mathbf{x}_{bg}) = \prod_{f=1}^{F} \frac{1}{A^{n_f}},$$

Where A is the area covered by the image.

[0040] Statistical learning is then used to estimate parameters of the statistical object class. This may be done using expectation maximization. The joint model probability density is estimated from the training set at 420. A probabilistic attempt is carried out to maximize the likelihood of the observed data, using expectation maximization (EM) to attempt to determine the maximum likelihood solution.

$$L(X^o|\tilde{\theta}) = \sum_{r=1}^{T} \log \sum_{\mathbf{h}_r} \sum_{\mathbf{b}_r} \sum_{\mathbf{n}_r} \int p(X_r^o, \mathbf{x}_r^m, \mathbf{h}_r, \mathbf{n}_r, \mathbf{b}_r|\theta)\, d\mathbf{x}_r^m,$$

Where θ represents the set of all parameters of the model.

This may be simplified as

$$Q(\bar{\theta}|\theta) = \sum_{\tau=1}^{T} E[\log p(X_\tau^o, \mathbf{x}_\tau^m, \mathbf{h}_\tau, \mathbf{n}_\tau, \mathbf{b}_\tau|\bar{\theta})].$$

Where E[.] denotes taking the expectation with respect to $p(\mathbf{h}_\tau, \mathbf{x}_\tau^m, \mathbf{n}_\tau, \mathbf{b}_\tau|X_\tau^o, \theta)$. As notation, the tilde implies that the values from a previous iteration are substituted. By using the EM technique, a local maximum may be found to thereby determine the maximum values.

[0041] At 130, update rules are determined. This may be done by decomposing Q into four parts:

$$\begin{aligned} Q(\tilde{\theta}|\theta) &= Q_1(\tilde{\theta}|\theta) + Q_2(\tilde{\theta}|\theta) + Q_3(\tilde{\theta}|\theta) + Q_4(\theta) \\ &= \sum_{\tau=1}^{T} E[\log p(\mathbf{n}_\tau|\theta)] + \sum_{\tau=1}^{T} E[\log p(\mathbf{b}_\tau|\theta)] \\ &+ \sum_{\tau=1}^{T} E[\log p(X_\tau^o, \mathbf{x}_\tau^m|\mathbf{h}_\tau, \mathbf{n}_\tau, \theta)] \\ &+ \sum_{\tau=1}^{T} E[\log p(\mathbf{h}_\tau|\mathbf{n}_\tau, \mathbf{b}_\tau)] \end{aligned}$$

The first three terms contain the parameters that will be updated while the last term includes no new parameters. First, the update rules for μ. Q3 depends only on μ tilde. Therefore, taking the derivative of the expected likelihood yields

$$\frac{\partial}{\partial \bar{\mu}} Q_3(\tilde{\theta}|\theta) = \sum_{\tau=1}^{T} E\left[\bar{\Sigma}^{-1}(\mathbf{z}_\tau - \bar{\mu})\right],$$

Where $\mathbf{z}^T = (\mathbf{x}^o\ \mathbf{x}^m)$ according to the definition above. Setting the derivative to 0 yields the update rule

$$\bar{\mu} = \frac{1}{T}\sum_{\tau=1}^{T} E[\mathbf{z}_\tau].$$

[0042] Next the update rule for Σ operates in an analogous way. The derivative with respect to the covariance matrix

$$\frac{\partial}{\partial \bar{\Sigma}^{-1}} Q_3(\tilde{\theta}|\theta) = \sum_{\tau=1}^{T} E\left[\frac{1}{2}\bar{\Sigma} - \frac{1}{2}(\mathbf{z}_\tau - \bar{\mu})(\mathbf{z}_\tau - \bar{\mu})^T\right].$$

Equating with zero leads to

$$\bar{\Sigma} = \frac{1}{T}\sum_{\tau=1}^{T} E[(\mathbf{z}_\tau - \bar{\mu})(\mathbf{z}_\tau - \bar{\mu})^T] = \frac{1}{T}\sum_{\tau=1}^{T} E[\mathbf{z}_\tau \mathbf{z}_\tau^T] - \bar{\mu}\bar{\mu}^T.$$

[0043] The update rule for p(b) may require considering Q_2 since this is the only term the depends on the parameters. The derivative with respect to p(b) yields

$$\frac{\partial}{\partial \bar{p}(\mathbf{b})} Q_2(\tilde{\theta}|\theta) = \sum_{\tau=1}^{T} \frac{E[\delta_{\mathbf{b},\mathbf{b}}]}{\bar{p}(\mathbf{b})}$$

And imposing the constraint

$$\sum_{\mathbf{b}\in B} \bar{p}(\mathbf{b}) = 1,$$

E.g. by adding a Lagrange multiplier term provides

$$\bar{p}(\bar{b}) = \frac{1}{T}\sum_{\tau=1}^{T} E[\delta_{b,\bar{b}}].$$

[0044] The update rule for M only depends on Q3, and hence differentiating this with respect to M yields

$$\frac{\partial}{\partial \bar{M}} Q_3(\bar{\theta}|\theta) = \sum_{\tau=1}^{T} \frac{E[n_\tau]}{\bar{M}} - I.$$

Equating to zero gives the intuitive result,

$$\bar{M} = \frac{1}{T}\sum_{\tau=1}^{T} E[n_\tau].$$

[0045] At 140 the sufficient statistics are determined. The posterior density is given by

$$p(h_\tau, x_\tau^m, n_\tau, b_\tau | X_\tau^o, \theta) = \frac{p(h_\tau, x_\tau^m, n_\tau, b_\tau, X_\tau^o|\theta)}{\sum_{h_\tau \in \mathcal{H}_h} \sum_{b_\tau \in \mathcal{B}} \sum_{n_\tau=0}^{\infty} \int p(h_\tau, x_\tau^m, n_\tau, b_\tau, X_\tau^o|\theta)\, dx_\tau^m}$$

Which may be simplified by noticing that if the summations are carried out in the order

$$\sum_{h_\tau \in \mathcal{H}_h} \sum_{b_\tau \in \mathcal{B}} \sum_{n_\tau=0}^{\infty}$$

Then certain simplifications may be made.

[0046] This enables selecting a hypothesis that is consistent with the observed data.

[0047] A final operation assesses the performance of the model at 430.

[0048] After applying all the feature detectors to the training samples, a greedy configuration search may be used to explore different model configurations. In general, configurations with a few different features may be explored. The configuration which yields the smallest training error, that is the smallest probability of misclassification, may be selected. This may be also augmented by one feature trying again all possible types. The best of these augmented models may be retained for subsequent augmentation. The process can be continued until a criterion for model complexity is met. For example, if no further improvement in detection performance is obtained before the maximum number of features is reached, then further operations should be unnecessary.

[0049] An iterative process may start with a random selection of parts. At each iteration, a test is made to determine whether replacing one model part with a randomly selected part improves or worsens the model. The replacement part is capped when the performance improves, otherwise the process is stopped when no more improvements

are possible. This can be done after increasing the total number of parts to the model to determine if additional parts should be added. This may be done by iteratively trying different combinations of small numbers of parts. Each iteration allows the parameters of the underlying probable listed model to be estimated. The iteration continues until the final model is obtained.

[0050] As an example, a recognition experiment may be carried out on comic strips. In an embodiment, the system attempted to learn the letters E, T, H and L. Two of the learned models are shown in figures 5A and 5B which respectively represent the model for the letter B and the model for the letter T.

[0051] The above has described the model configuration being selected prior to the EM phase. However, this could conceivably require a model to be fit to each possible model configuration. This may be avoided by producing a more generic model.

[0052] This system has been used to identify handwritten letters e.g. among comic books, recognition of faces within images, representing the rear views of cars, letters, leaves and others.

[0053] This system may be used for a number of different applications. In a first application, the images may be

indexed into image databases. Images may be classified automatically enabling a user to search images that include given objects. For example, a user could show this system can image that includes a frog, and obtain back from at all images that included frogs.

[0054] Autonomous agents/vehicle/robots could be used. For example, this system could allow a robot to Rome and area and learn all the objects were certain objects are president. The vehicle could then report events that differ from the normal background or find certain things.

[0055] This system could be used for automated quality control, for example, this system could be shown a number of defective items, and find similar defective items. Similarly, the system could be used to train for dangerous situations.

[0056] Another application is in toys and entertainments e.g. a robotic device. Finally, visual screening in industries such as the biomedical industry in which quality control applications might be used.

[0057] Although only a few embodiments have been disclosed in detail above, other modifications are possible. All such modifications are intended to be encompassed within the following claims, in which: